Statistics: New Foundations, Toolbox, and Machine Learning Recipes

By Vincent Granville, Ph.D.

MLTechniques.com

July 2019

This book is intended for busy professionals working with data of any kind: engineers, BI analysts, statisticians, operations research, AI and machine learning professionals, economists, data scientists, biologists, and quants, ranging from beginners to executives. In about 300 pages, it covers many new topics, offering a fresh perspective on the subject, including rules of thumb and recipes that are easy to automate or integrate in black-box systems, as well as new model-free, data-driven foundations to statistical science and predictive analytics. The approach focuses on robust techniques; it is bottom-up (from applications to theory), in contrast to the traditional top-down approach. The material is accessible to practitioners with a one-year college-level exposure to statistics and probability. The compact and tutorial style, featuring many applications with numerous illustrations, is aimed at practitioners, researchers, and executives in various quantitative fields.

New ideas, advanced topics and state-of-the-art research are discussed in simple English, without using jargon or arcane theory. It unifies topics that are usually part of different fields (machine learning, statistics, computer science), broadening the knowledge and interest of the reader in ways that are not found in any other book. This short book contains a large amount of condensed material that would typically be covered in 1,000 pages in traditional publications, including data sets, source code, and Excel spreadsheets. Thanks to cross-references and redundancy, the chapters can be read independently, in random order.

This book is based on several core articles and many tutorials that I have written over the last few years. Chapters are organized and grouped by themes: natural language processing (NLP), resampling, time series, central limit theorem, statistical tests, boosted models (ensemble methods), tricks and special topics, appendices, and so on. It is available for Data Science Central members exclusively. The text in blue consists of clickable links to provide the reader with additional references. Source code and Excel spreadsheets summarizing computations, are also accessible as hyperlinks for easy copy-and-paste or replication purposes.

About the author

Vincent Granville is a start-up entrepreneur, patent owner, author, investor, pioneering data scientist with 30 years of corporate experience in companies small and large (eBay, Microsoft, NBC, Wells Fargo, Visa, CNET) and a former VC-funded executive, with a strong academic and research background including Cambridge University.

Content

Part 1 - Machine Learning Fundamentals and NLP

We introduce a simple ensemble technique (or boosted algorithm) known as *Hidden Decision Trees*, combining robust regression with unusual decision trees, useful in the context of transaction scoring. We then describe other original and related machine learning techniques for clustering large data sets, structuring unstructured data via indexation (a natural language processing or NLP technique), and perform feature selection, with Python code and even an Excel implementation.

Part 2 - Applied Probability and Statistical Science

We discuss traditional statistical tests to detect departure from randomness (the null hypothesis) with applications to sequences (the observations) that behave like stochastic processes. The central limit theorem (CLT) is revisited and generalized with applications to time series (both univariate and multivariate) and Brownian motions. We discuss how weighted sums of random variables and stable distributions are related to the CLT, and then explore mixture models -- a better framework to represent a rich class of phenomena. Applications are numerous, including optimum binning for instance. The last chapter summarizes many of the statistical tests used earlier.

Part 3 - New Foundations of Statistical Science

We set the foundations for a new type of statistical methodology fit for modern machine learning problems, based on generalized resampling. Applications are numerous, ranging from optimizing cross-validation to computing confidence intervals, without using classic statistical theory, *p*-values, or probability distributions. Yet we introduce a few new fundamental theorems, including one regarding the asymptotic properties of generic, model-free confidence intervals.

Part 4 - Case Studies, Business Applications

These chapters deal with real life business applications. Chapter 18 is peculiar in the sense that it features a very original business application (in gaming) described in details with all its components, based on the material from the previous chapters. Then we move to more traditional machine learning use cases. Emphasis is on providing sound business advice to data science managers and executives, by showing how data science can be successfully leveraged to solve problems. The presentation style is compact, focusing on strategy rather than technicalities.

Part 5 - Additional Topics

Here we cover a large number of topics, including sample size problems, automated exploratory data analysis, extreme events, outliers, detecting the number of clusters, *p*-values, random walks, scale-invariant methods, feature selection, growth models, visualizations, density estimation, Markov chains, A/B testing, polynomial regression, strong correlation and causation, stochastic geometry, *K* nearest neighbors, and even the exact value of an intriguing integral computed using statistical science, just to name a few.